

WHY RANDOM RESHUFFLING BEATS STOCHASTIC GRADIENT DESCENT

M. GÜRBÜZBALABAN*, A. OZDAGLAR* , AND P.A. PARRILO*

Abstract. We analyze the convergence rate of the random reshuffling (RR) method, which is a randomized first-order incremental algorithm for minimizing a finite sum of convex component functions. RR proceeds in cycles, picking a uniformly random order (permutation) and processing the component functions one at a time according to this order, i.e., at each cycle, each component function is sampled without replacement from the collection. Though RR has been numerically observed to outperform its with-replacement counterpart stochastic gradient descent (SGD), characterization of its convergence rate has been a long standing open question. In this paper, we answer this question by showing that when the component functions are quadratics or smooth and the sum function is strongly convex, RR with iterate averaging and a diminishing stepsize $\alpha_k = \Theta(1/k^s)$ for $s \in (1/2, 1)$ converges at rate $\Theta(1/k^{2s})$ with probability one in the suboptimality of the objective value, thus improving upon the $\Omega(1/k)$ rate of SGD. Our analysis draws on the theory of Polyak-Ruppert averaging and relies on decoupling the dependent cycle gradient error into an independent term over cycles and another term dominated by α_k^2 . This allows us to apply law of large numbers to an appropriately weighted version of the cycle gradient errors, where the weights depend on the stepsize. We also provide high probability convergence rate estimates that shows decay rate of different terms and allows us to propose a modification of RR with convergence rate $\mathcal{O}(\frac{1}{k^2})$.

1. Introduction: First-order incremental methods. We consider the following unconstrained optimization problem where the objective function is the sum of a large number of component functions:

$$\min f(x) := \sum_{i=1}^m f_i(x) \quad \text{s.t.} \quad x \in \mathbb{R}^n \quad (1.1)$$

with each $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ convex. Such a problem arises in many contexts and applications including regression or more generally parameter estimation problems (where $f_i(x)$ is the loss function representing the error between the output and the prediction of a parametric model) [5,12] [2,3], minimization of an expected value of a function (where the expectation is taken over a finite probability distribution or approximated by an m -sample average) [10,31], machine learning [31,33,34], or distributed optimization over networks [22,23,26].

One widely studied approach for solving problem (1.1) is the *deterministic incremental gradient (IG) method* [4–6]). IG method is similar to the standard gradient method with the key difference that at each iteration, the decision vector is updated incrementally by taking sequential steps along the gradient of the component functions f_i in a cyclic order. Hence, we can view each outer iteration k as a cycle of m inner iterations: starting from

*Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. email: {mertg, asuman,parrilo}@mit.edu.

initial point $x_0^0 \in \mathbb{R}^n$, for each $k \geq 0$, we update the iterate x_i^k as

$$x_i^k := x_{i-1}^k - \alpha_k \nabla f_i(x_{i-1}^k), \quad i = 1, 2, \dots, m, \quad (1.2)$$

where $\alpha_k > 0$ is a stepsize with the convention that $x_0^{k+1} = x_m^k$.

Intuitively, it is clear that slow progress can be obtained if the functions that are processed consecutively have small gradients in a certain region. Indeed, the performance of IG is known to be sensitive to the order functions are processed [6, Example 2.1.3]. If there is a favorable order σ (defined as a permutation of $\{1, 2, \dots, m\}$) that can be obtained by exploiting the problem-specific knowledge, then it is advisable to process the functions with this order instead with the iterations:

$$x_i^k := x_{i-1}^k - \alpha_k \nabla f_{\sigma(i)}(x_{i-1}^k), \quad i = 1, 2, \dots, m. \quad (1.3)$$

However, in general a favorable order is not known in advance, and a common approach is choosing the indices of functions to process as independent and uniformly distributed samples from the set $\{1, 2, \dots, m\}$. This way no particular order is favored, making the method less vulnerable to particularly bad orders. This approach amounts to sampling the order *with replacement* from the set of indices $\{1, 2, \dots, m\}$ and is called the *Stochastic Gradient Descent* (SGD) method, a.k.a. *Robbins-Monro* algorithm [30]. SGD is strongly related to the classical field of stochastic approximation [20]. Recently it has received a lot of attention due to its applicability to large-scale problems and became popular especially in machine learning applications (see e.g. [8–10, 36]).

An alternative popular approach that works well in practice is following an approach between SGD and IG, sampling the functions randomly but not allowing repetitions, that is sampling the component functions at each cycle *without-replacement*. Specifically, at each cycle k , we draw a permutation σ_k of $\{1, 2, \dots, m\}$ randomly from the set

$$\Gamma = \{\sigma : \sigma \text{ is a permutation of } \{1, 2, \dots, m\}\}$$

and process the functions with this order

$$x_i^k := x_{i-1}^k - \alpha_k \nabla f_{\sigma_k(i)}(x_{i-1}^k), \quad i = 1, 2, \dots, m, \quad (1.4)$$

where $\alpha_k > 0$ is a stepsize. A key point is that the permutations σ_k are i.i.d. and uniformly distributed over Γ . We set $x_0^{k+1} = x_m^k$ as before and refer to $\{x_0^k\}$ as the *outer iterates*. This method is called the *Random Reshuffling* (RR) method ([6, Section 2.1]).

2. Motivation and summary of contributions. Without-replacement sampling schemes are often easier to implement efficiently compared to with-replacement sampling schemes, guarantee that every item in the data set is touched at least once, and often have

better empirical performance than their with-replacement counterparts [7, 9, 11, 16, 27, 28]. For instance, Bottou [7] compares SGD and RR methods and found that RR converges with a rate close to $\sim 1/k^2$ whereas SGD is much slower hitting its min-max lower bound of $\Omega(1/k)$ for strongly convex objective functions [35], [1]. Many other papers listed above report a similar behavior. This discrepancy in rate between RR and SGD is not only observed for large m but also for small m , and understanding it theoretically has been a long-standing open problem [4, 28], which will be the subject of this work.

To our knowledge, the only existing theoretical analysis for RR is given by a recent paper of Recht and Ré [27] who focus on least mean squares optimization and formulate a non-commutative arithmetic-geometric mean inequality conjecture that would prove that the expected convergence rate of RR is faster than that of SGD. The conjecture that needs to be established involves products of an arbitrary number N of matrices for every positive integer N . This conjecture is still open but has been proven in some special cases (for $N = 2$ [27], for $N = 3$ [19] and when N is a multiple of 3 [37]). Recht and Ré also analyzed a special case of (1.1) (that arises when $f_i(x) = (a_i^T x - y_i)^2$ is a quadratic function where a_i is a column vector that is randomly generated according to a random model and y_i is a scalar) and show that the expected without-sampling rate is better than the expected with-sampling rate with high probability (probability with respect to the random data generation model). Despite these advances, there has been a lack of convergence theory for RR that explains its fast convergence. Analyzing algorithms based on without-replacement sampling such as RR is more difficult than with-replacement based approaches such as SGD. The reason is that the underlying independence assumption for with-replacement sampling allows an elegant analysis with classical martingale convergence theory [21, 24], whereas the iterates in without-replacement sampling are correlated and therefore martingale convergence theorems are not directly applicable. The aim of our paper is to fill this theoretical gap for the case when the objective function f in (1.1) is strongly convex. We now summarize our contributions.

Building on the recent convergence rate results for the cyclic IG in [18], we first present a key result that provides an upper bound for the distance from optimal solution of the iterates generated by an incremental method that processes component functions with an *arbitrary fixed order* and uses a stepsize $\Theta(1/k^s)$ for $s \in (0, 1)$. This upper bound decays at rate $\mathcal{O}(1/k^s)$ and depends on the strong convexity constant of the sum function and an order dependent parameter, which combines Hessian matrices and gradients of the component functions at the optimal solution according to the given order.

We then focus on the case when the component functions are quadratics which corresponds to the least squares minimization. We consider the q -suffix averages of the iterates

generated by RR for some $q \in (0, 1]$ (which is obtained by averaging the last qk iterates at iteration k) and show that with a stepsize $\Theta(1/k^s)$ for $s \in (1/2, 1)$, they converge *almost surely at rate $\mathcal{O}(1/k^s)$ to the optimal solution*. We provide an explicit characterization of the rate constant in terms of the averaging constant q and the almost sure limit of the average gradient error normalized by the average stepsize. Using strong convexity of the objective function, this implies an almost sure convergence at rate $\Theta(1/k^{2s})$ in the suboptimality of the objective value. The analysis of RR is complicated by the fact that the cumulative gradient error over cycles are dependent. A key step in our proof is to decouple the cycle gradient error into a term independent over cycles and another term that scales as $\mathcal{O}(\alpha_k^2)$. This allows us to use law of large numbers for a properly weighted average of the cycle error gradient sequence (where the weights depend on the stepsize) and show almost sure convergence of the q -suffix averaged iterates. Another key component of our analysis is to adapt the Polyak-Ruppert averaging techniques developed for SGD [21, 24] to RR.

We next provide a high probability convergence rate estimate for the distance of q -suffix averages to the optimal solution that consists of two terms, with the first term corresponding to a $1/k^s$ decay of a “bias” term (where bias is defined as the expected value of the q -suffix averaged cycle gradient errors ignoring the second-order corrections on the order of $1/k$) and the second term representing a $1/k$ decay for $0 < q < 1$ (and $\log k/k$ decay for $q = 1$). We use the characterization of the bias to approximate it with a term that can be computed over the last cycle in a given fixed number of iterations. Removing the bias term enables a modification of RR with convergence rate $\mathcal{O}(1/k^2)$. These results are based on martingale concentration techniques.

Finally, we show that our results extend to the more general case when component functions are smooth (twice continuously differentiable) under a Lipschitz assumption on the Hessian, which allows us to control the second order term in a Taylor expansion of the gradient.

Outline: The outline of the paper is as follows. In the next section, we introduce our definitions and assumptions. Section 4 focuses on the case when component functions are quadratics. We first present a convergence rate estimate for IG with a fixed arbitrary order. We then focus on RR and study convergence of averaged iterates to the optimal solution. Section 5 extends our results to smooth functions. Section 6 proposes a new algorithm that can accelerate RR. Finally, we conclude with a summary of our work in Section 7. Some of the technical lemmas required in the details of the proofs are deferred to Sections A and B of the Appendix.

Notation: We study the pointwise dominance of stochastic sequences by deterministic sequences and use the following notation. Let $x_k = x_k(\omega)$ be a stochastic real-valued

sequence (where ω can be thought as the source of randomness) and y_k be a real-valued deterministic sequence. We write

$$x_k = \mathcal{O}(y_k) \iff \exists h > 0, \exists k_0 \text{ such that } |x_k| \leq h|y_k| \quad \forall k \geq k_0, \forall \omega,$$

where h and k_0 are independent of ω (Note that the requirement is that this inequality holds for all ω , not just for almost all ω). Similarly, given another deterministic sequence z_k , we introduce the inequality version of this definition:

$$x_k \leq y_k + o(z_k) \iff \forall \varepsilon > 0, \exists k_0(\varepsilon) \text{ such that } z_k^{-1}|x_k(\omega) - y_k| \leq \varepsilon, \quad \forall k \geq k_0(\varepsilon), \forall \omega$$

where k_0 depends on ε but is independent of ω . When x_k is deterministic, these definitions reduce to the standard definitions of $\mathcal{O}(\cdot)$ and $o(\cdot)$ for deterministic sequences. For random x_k , the only difference is that we require the constants to be independent of the choice of ω . For example, if x_k is uniformly distributed over $[0, 10]$, we write $x_k = \mathcal{O}(1)$.

Throughout the paper, $\|\cdot\|$ denotes the vector or matrix 2-norm (maximum singular value).

3. Preliminaries. We first rewrite the outer RR iterations (1.4) as

$$\frac{x_0^k - x_0^{k+1}}{\alpha^k} = \nabla f(x_0^k) + E_k \tag{3.1}$$

where

$$E_k = \sum_{i=1}^m \left(\nabla f_{\sigma_k(i)}(x_{i-1}^k) - \nabla f_{\sigma_k(i)}(x_0^k) \right) \tag{3.2}$$

can be viewed as the cumulative gradient errors associated with the cycle k . When the iterates are averaged over time, the random gradient errors E_k will be averaged too. The key idea behind our rate result is to show a limit theorem for a weighted average of the E_k sequence in a sense we will make precise.

The *average* of the outer iterate sequence is given by

$$\bar{x}_k := \frac{\sum_{j=0}^{k-1} x_0^j}{k}.$$

It is well known that computing this (moving) average can be done efficiently in a dynamic manner by storing only a vector of length n . We also consider averaging only the most recent iterates, i.e. at iteration k , averaging the last qk iterates for some constant $q \in (0, 1]$:

$$\bar{x}_{q,k} := \frac{\sum_{j=(1-q)k}^{k-1} x_0^j}{qk}, \quad 0 < q \leq 1.$$

The generated sequence is referred to as the q -*suffix average* of the sequence x_0^k . For SGD, it has been shown that q -suffix averaging with $0 < q < 1$ leads to better performance than

averaging (which corresponds to the $q = 1$ case by definition), improving the convergence rate in the suboptimality of the function value from $\log k/k$ to $1/k$ [25, 32]. This is inline with our results in Section 4 which show faster rate for the $0 < q < 1$ case. The parameter q can be thought as a measure of how much memory one uses during the averaging process. We define the q -suffix average of the stepsize in a similar way:

$$\bar{\alpha}_{q,k} = \frac{\sum_{j=(1-q)k}^{k-1} \alpha_j}{qk}, \quad 0 < q \leq 1.$$

We assume that the sum function f is strongly convex. Such functions arise naturally in support vector machines and other regularized learning algorithms or regression problems (see e.g. [25, 29, 31]).

ASSUMPTION 3.1. *The sum function $f(x) = \sum_{i=1}^m f_i(x)$ is strongly convex, i.e., there exists a constant $c > 0$ such that the function $f(x) - \frac{c}{2}\|x\|^2$ is convex on \mathbb{R}^n .*

A consequence of this assumption is that there exists a unique optimal solution to (1.1) which we denote by x^* . Another consequence is that the Hessian at the optimal solution is invertible since

$$H_* := \nabla^2 f(x^*) \succeq cI_n \succ 0 \quad (3.3)$$

where I_n is the $n \times n$ identity matrix. We will state our rate results in terms of the distance of the outer iterate at step k to the optimal solution which we denote by

$$\text{dist}_k = \|x_0^k - x^*\|.$$

We start with analyzing the case when the component functions are quadratics. This corresponds to least mean squares optimization which arises frequently in applications. Then, we extend our theory to general *smooth* (twice continuously differentiable) component functions.

4. Quadratic component functions. Let $f_i(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ be a quadratic function of the form

$$f_i(x) = \frac{1}{2}x_i^T P_i x - q_i^T x + r_i, \quad i = 1, 2, \dots, m, \quad (4.1)$$

where P_i is a symmetric $n \times n$ square matrix, $q_i \in \mathbb{R}^n$ is a column vector and r_i is a scalar. Note that f_i has Lipschitz gradients, i.e.,

$$\|\nabla f_i(y) - \nabla f_i(z)\| \leq L_i \|y - z\|, \quad \forall y, z \in \mathbb{R}^n,$$

where $L_i = \|P_i\|$. It follows from the triangular inequality that f has Lipschitz gradients with Lipschitz constant

$$L = \sum_{i=1}^m L_i. \quad (4.2)$$

In the classical stochastic approximation theory, the standard approach is to assume that the gradient errors are either i.i.d. or that they form a martingale difference sequence and then apply martingale central limit theorems [13, 15, 20]. However, for analyzing the RR iterations, one cannot make such an assumption due to the dependencies between the sampled indices $\sigma_k(i)$ and $\sigma_k(j)$ for $i \neq j$ within a cycle. For instance, it can be shown that E_k is not a martingale difference sequence with respect to the standard filtration \mathcal{F}_k that contains all the past information till the beginning of the cycle k . Gradient errors E_{k_1} and E_{k_2} are also dependent for any $k_1 \neq k_2$ as they both depend on the history of the iterates and the sampled indices where standard limit theorems would not be directly applicable. There is some literature that analyzes SGD under correlated noise [20, Ch. 6], but the noise needs to have a special structure (such as a mixing property) which does not seem to be applicable to the analysis of RR.

Our approach is to decouple the sequence E_k , representing it as a sum of an i.i.d. term (that has a non-zero expectation) and a perturbation term that gets smaller and becomes negligible as the iterates approach the optimal solution. Characterizing the decay rate of this perturbation builds on a global convergence result for RR which is the subject of the next section.

4.1. Global convergence. In order to establish global convergence of the RR method, we make use of some recent convergence results developed for the (deterministic) IG method. The following theorem from [18] provides an upper bound for the convergence rate of an incremental gradient method which processes the component functions with a fixed permutation σ of $\{1, 2, \dots, m\}$.

THEOREM 4.1. [18] *Let $f_i(x)$ be a quadratic function of the form,*

$$f_i(x) = \frac{1}{2}x_i^T P_i x - q_i^T x + r_i$$

where P_i is a symmetric $n \times n$ square matrix, $q_i \in \mathbb{R}^n$ is a column vector and r_i is a scalar for $i = 1, 2, \dots, m$. Suppose Assumption 3.1 holds. Consider the iterates $\{x_0^k\}$ generated by the iterations (1.4) with a fixed order σ and stepsize $\alpha_k = R/(k+1)^s$ where $R > 0$ and $s \in (0, 1)$. Then¹,

$$\text{dist}_k \leq \frac{RM_\sigma}{c} \frac{1}{k^s} + o\left(\frac{1}{k^s}\right) \quad \text{where} \quad M_\sigma = \left\| \sum_{1 \leq i < j \leq m} P_{\sigma(j)} \nabla f_{\sigma(i)}(x^*) \right\|.$$

¹The original result was stated for $\sigma = \{1, 2, \dots, m\}$ but here we translate this result into an arbitrary permutation σ of $\{1, 2, \dots, m\}$ by noting that processing the set of functions $\{f_1, f_2, \dots, f_m\}$ with order σ is equivalent to processing the permuted functions $\{f_{\sigma_1}, f_{\sigma_2}, \dots, f_{\sigma_m}\}$ with order $\{1, 2, \dots, m\}$.

This theorem provides an $\mathcal{O}(\frac{1}{k^s})$ upper bound on the rate where the rate constant M_σ depends on the order σ . Defining

$$G_* = \sup_{1 \leq i \leq m} \|\nabla f_i(x^*)\|, \quad (4.3)$$

as $L_i = \|P_i\|$ for each i , it follows from the triangle inequality

$$M_\sigma \leq M_\Gamma := \sup_{\sigma \in \Gamma} M_\sigma \leq \sup_{\sigma \in \Gamma} \sum_{1 \leq i < j \leq m} L_{\sigma(j)} G_* \leq LmG_* \quad (4.4)$$

where L is the Lipschitz constant of the gradient of f defined by (4.2). By replacing M_σ by M_Γ in Theorem 4.1 one can get an upper bound on the worst-case convergence rate that applies to any choice of fixed order σ . Using a similar argument along the lines of the proof of Theorem 4.1, it is straightforward to show that RR does never perform any slower than this worst-case convergence rate which is the subject of the next result. The idea is to bound the stochastic dist_k sequence from above pointwise. The proof is a simple exercise and is omitted.

COROLLARY 4.2. *Under the setting of Theorem 4.1, if σ is sampled uniformly at each cycle instead of being kept fixed, then*

$$\text{dist}_k \leq \frac{RM_\Gamma}{c} \frac{1}{k^s} + o\left(\frac{1}{k^s}\right)$$

where M_Γ is deterministic and is defined by (4.4).

Although Corollary 4.2 provides a simple worst-case upper bound on the rate, it will be a powerful tool for analyzing the gradient error process and proving our main theorem in the next section which establishes a much stronger convergence rate result for the averaged iterates.

4.2. Convergence rate with averaging for quadratics. The following theorem characterizes the rate of convergence of the averages of iterates generated by RR. Part (i) of this theorem shows that q -suffix averages of the RR iterates converge at rate $1/k^s$ to the optimal solution almost surely with a stepsize $\Theta(1/k^s)$ for $s \in (1/2, 1)$. By strong convexity, this translates into a rate of $\Theta(1/k^{2s})$ for the suboptimality of the objective value. The result is based on decoupling the cycle gradient errors E_k into a $\Theta(\alpha_k)$ term independent over the cycles and another $\mathcal{O}(\alpha_k^2)$ term that becomes negligible in the limit. Part (ii) is a high-probability convergence rate estimate for the approximation error $\bar{x}_{q,k} - x^*$. The approximation error consists of two terms, the first term $b_{q,k}$ which we call the “bias” term is deterministic and decays like $1/k^s$. It comes from the expected value of the independent part of the gradient cycle errors which may be different than zero. The second part is on the order of $1/k$ for $0 < q < 1$ (and $\log k/k$ when $q = 1$) and it is based on the Azuma-Hoeffding inequality for martingale concentration. Finally, part (iii) is on estimating the

bias term $b_{q,k}$ with another quantity $\hat{b}_{q,k}$. It shows that by subtracting the estimated bias from the averaged iterates, we can approximate the optimal solution up to an $\mathcal{O}(1/k)$ error in distances or equivalently up to an $\mathcal{O}(1/k^2)$ error in the suboptimality of the objective function. In Section 6, this result will be fundamental for Algorithm 6 that accelerates the convergence of RR from $\Theta(1/k^{2s})$ to $\mathcal{O}(1/k^2)$ with high probability in the suboptimality of the function value.

THEOREM 4.3. *Let $f_i(x)$ be a quadratic function of the form*

$$f_i(x) = \frac{1}{2}x_i^T P_i x - q_i^T x + r_i$$

where P_i is a symmetric $n \times n$ square matrix, $q_i \in \mathbb{R}^n$ is a column vector and r_i is a scalar for $i = 1, 2, \dots, m$. Consider the q -suffix averages $\bar{x}_{q,k}$ of the RR iterates generated by the iterations (1.4) with stepsize $\alpha_k = \frac{R}{(k+1)^s}$ where $R > 0$ and $s \in (\frac{1}{2}, 1)$. Suppose that Assumption 3.1 holds. Then the following statements are true:

(i) For any $0 < q \leq 1$,

$$\lim_{k \rightarrow \infty} k^s (\bar{x}_{q,k} - x^*) = a_q(s) H_*^{-1} \mu_* \quad a.s.$$

where $H_* = \sum_{i=1}^m P_i$ is the Hessian matrix at the optimal solution,

$$a_q(s) = -\frac{1 - (1-q)^{1-s}}{q} R \quad \text{and} \quad \mu_* = \frac{1}{2} \sum_{i=1}^m P_i \nabla f_i(x^*). \quad (4.5)$$

(ii) With probability at least $1 - \delta$, we have

$$\bar{x}_{q,k} - x^* = b_{q,k} + \mathcal{O}\left(\frac{\sqrt{\log(1/\delta)}}{k}\right) + \begin{cases} \mathcal{O}\left(\frac{\log k}{k}\right) & \text{if } q = 1 \\ \mathcal{O}\left(\frac{1}{k}\right) & \text{if } 0 < q < 1, \end{cases}$$

where

$$b_{q,k} = -\bar{\alpha}_{q,k} H_*^{-1} \mu_* \quad (4.6)$$

is deterministic. The constants hidden by $\mathcal{O}(\cdot)$ depend only on G_* , L , m , R , c , q and s .

(iii) Let

$$\hat{b}_{q,k} = -\bar{\alpha}_{q,k} \left[\sum_{i=1}^m P_{\sigma_k(i)} \right]^{-1} \sum_{i=1}^m P_{\sigma_k(i)} \nabla f_{\sigma_k(i)}(x_{i-1}^k). \quad (4.7)$$

Then,

$$\hat{b}_{q,k} = b_{q,k} + \mathcal{O}(\alpha_k^2).$$

It follows from part (ii) that with probability at least $1 - \delta$,

$$(\bar{x}_{q,k} - \hat{b}_{q,k}) - x^* = \mathcal{O}\left(\frac{\sqrt{\log(1/\delta)}}{k}\right) + \begin{cases} \mathcal{O}\left(\frac{\log k}{k}\right) & \text{if } q = 1 \\ \mathcal{O}\left(\frac{1}{k}\right) & \text{if } 0 < q < 1. \end{cases}$$

Proof.

(i) Taking the q -suffix averages of both sides of Equation (3.1), we obtain

$$I_{q,k} := \frac{\sum_{j=(1-q)k}^{k-1} (x_0^j - x_0^{j+1}) \alpha_j^{-1}}{qk} = \frac{\sum_{j=(1-q)k}^{k-1} \nabla f(x_0^j) + E_j}{qk}. \quad (4.8)$$

As f is a quadratic, the first order Taylor series for the gradient of f is exact:

$$\nabla f(x_0^j) = H_*(x_0^j - x^*). \quad (4.9)$$

Therefore, (4.8) becomes

$$I_{q,k} := \frac{\sum_{j=(1-q)k}^{k-1} H_*(x_0^j - x^*) + E_j}{qk}$$

which is equivalent to

$$I_{q,k} = H_*(\bar{x}_{q,k} - x^*) + \frac{\sum_{j=(1-q)k}^{k-1} E_j}{qk} = H_*(\bar{x}_{q,k} - x^*) + \bar{\alpha}_{q,k} Y_{q,k} \quad (4.10)$$

where $Y_{q,k}$ is defined as

$$Y_{q,k} := \frac{1}{\bar{\alpha}_{q,k}} \frac{\sum_{j=(1-q)k}^{k-1} E_j}{qk} = \frac{\sum_{j=(1-q)k}^{k-1} E_j}{\sum_{j=(1-q)k}^{k-1} \alpha_j} \quad (4.11)$$

and can be interpreted as the (q -suffix) averaged gradient error sequence E_j normalized by the (q -suffix) averaged stepsize sequence α_j . Since H_* is invertible by the strong convexity of f (see (3.3)), we can rewrite (4.10) as

$$\bar{x}_{q,k} - x^* = -H_*^{-1} \bar{\alpha}_{q,k} Y_{q,k} + H_*^{-1} I_{q,k} \quad (4.12)$$

$$= -H_*^{-1} \bar{\alpha}_{q,k} Y_{q,k} + \begin{cases} \mathcal{O}(\frac{1}{k}) & \text{if } 0 < q < 1 \\ \mathcal{O}(\frac{\log k}{k}) & \text{if } q = 1. \end{cases} \quad (4.13)$$

where we used the inequality $\|H_*^{-1}\| \leq 1/c$ implied by Equation (3.3) and Lemma A.2 from the appendix to provide an upper bound for the second term in the first equality. Note that, as a consequence of Lemma A.2, $\mathcal{O}(\cdot)$ notation above hides a constant that depends only on the parameters G_*, L, c, m, R, s, q and also dist_0 when $q = 1$. As the stepsize sequence is monotonically decreasing, we have the bounds

$$\int_{(1-q)k}^k \frac{R}{(x+2)^s} dx \leq \sum_{j=(1-q)k}^k \alpha_j = \sum_{j=(1-q)k}^k \frac{R}{(k+1)^s} \leq 1 + \int_{(1-q)k}^k \frac{R}{(x+1)^s} dx.$$

Dividing each term by qk , after a straightforward integration we obtain

$$\bar{\alpha}_{q,k} = \frac{(k+1)^{1-s} - ((1-q)k+1)^{1-s} + \mathcal{O}(1)}{qk} R = -\frac{a_q(s)}{k^s} + \mathcal{O}(\frac{1}{k}).$$

Then, multiplying both sides of (4.13) by k^s , taking limits and using the fact that $Y_{q,k} \rightarrow \mu_*$ a.s. from Lemma A.3, we obtain the claimed result.

(ii) By parts (i) and (iii) of Lemma A.3 from the appendix that relates the gradient error sequence E_j to a sequence of i.i.d. variables $v(\sigma_j)$, for $0 < q \leq 1$,

$$Y_{q,k} = \frac{\sum_{j=(1-q)k}^{k-1} E_j}{\sum_{j=(1-q)k}^{k-1} \alpha_j} = \frac{\sum_{j=(1-q)k}^{k-1} \alpha_j v(\sigma_j) + \mathcal{O}(\alpha_j^2)}{\sum_{j=(1-q)k}^{k-1} \alpha_j}. \quad (4.14)$$

We first give a proof for $q = 1$, the proof for the remaining $q \in (0, 1)$ case will be similar. Assume $q = 1$. Plugging $q = 1$ and Equation (4.14) into (4.13), we obtain

$$\begin{aligned} \bar{x}_{1,k} - x^* &= \mathcal{O}\left(\frac{\log k}{k}\right) - H_*^{-1} \bar{\alpha}_{1,k} Y_{1,k} \\ &= \mathcal{O}\left(\frac{\log k}{k}\right) - H_*^{-1} \left(\frac{\sum_{j=0}^{k-1} \alpha_j (v(\sigma_j) - \mu_*)}{k} + \frac{\sum_{j=0}^{k-1} \alpha_j \mu_* + \mathcal{O}(\alpha_j^2)}{k} \right) \\ &= b_{1,k} + \mathcal{O}\left(\frac{\log k}{k}\right) - H_*^{-1} \frac{\sum_{j=0}^{k-1} \alpha_j (v(\sigma_j) - \mu_*)}{k} - H_*^{-1} \sum_{j=0}^{k-1} \frac{\mathcal{O}(\alpha_j^2)}{k} \\ &= b_{1,k} + \mathcal{O}\left(\frac{\log k}{k}\right) - H_*^{-1} \frac{\sum_{j=0}^{k-1} \alpha_j (v(\sigma_j) - \mu_*)}{k} \end{aligned} \quad (4.15)$$

where $b_{1,k}$ is defined by (4.6) and we used in the last step the fact that for $s > 1/2$

$$\sum_{j=0}^{\infty} \alpha_j^2 = \sum_{j=1}^{\infty} \frac{R^2}{j^{2s}} = R^2 \zeta(2s) < \infty \quad (4.16)$$

where $\zeta(\cdot)$ is the Riemann-Zeta function. We now study the asymptotic behavior of the last summation term in (4.15) by introducing the process

$$S_{1,k} = \sum_{j=0}^{k-1} Z_j, \quad Z_j := \alpha_j (v(\sigma_j) - \mu_*), \quad k \geq 0,$$

where the random variables Z_j/α_j are i.i.d. with the convention that $S_{1,0} = 0$. Equipped with this definition, (4.15) becomes

$$\bar{x}_{1,k} - x^* = b_{1,k} + \mathcal{O}\left(\frac{\log k}{k}\right) - H_*^{-1} \frac{S_{1,k}}{k}. \quad (4.17)$$

The random variables Z_j are independent, centered and have an identical distribution up to the scaling factor α_j . Therefore, $S_{1,k}$ is a sum of centered random variables satisfying:

$$\|S_{1,k} - S_{1,k-1}\| = \|Z_{k-1}\| = \|\alpha_{k-1} (v(\sigma_{k-1}) - \mu_*)\| \quad (4.18)$$

$$\leq \gamma_{k-1} := \alpha_{k-1} LmG_*. \quad (4.19)$$

where we used Equation (A.12) in the last inequality. Then, by the Azuma-Hoeffding inequality, for every $t > 0$,

$$\begin{aligned} \mathbb{P}\left(\left\|\frac{S_{1,k}}{k}\right\| > \frac{t}{k}\right) &\leq 2 \exp\left(-\frac{t^2}{2 \sum_{j=0}^{k-1} \gamma_j^2}\right) \\ &= 2 \exp\left(-\frac{t^2}{\beta}\right) \end{aligned}$$

where $\beta = 2 \sum_{j=0}^{\infty} \gamma_j^2 < \infty$ as α_j is square-summable (see (4.16)). Note that β depends only on G_*, L, m and the stepsize parameters R and s . It is easy to see that selecting $t \geq t_\delta = \sqrt{\beta \log(2/\delta)}$ makes the right-hand side $\leq \delta$. Therefore for any $\delta > 0$, with probability at least $1 - \delta$,

$$\left\| \frac{S_{1,k}}{k} \right\| \leq \frac{\sqrt{\beta \log(2/\delta)}}{k} \quad (4.20)$$

which if inserted into the expression (4.17) completes the proof for the $q = 1$ case. For $0 < q < 1$ case, the same line of reasoning applies except that we can improve the $\mathcal{O}(\log k/k)$ term in the expression (4.17) to $\mathcal{O}(1/k)$, this is justified by (4.13). Then, this leads to

$$\bar{x}_{q,k} - x^* = b_{q,k} + \mathcal{O}\left(\frac{1}{k}\right) - H_*^{-1} \frac{S_{q,k}}{k} \quad (4.21)$$

where $S_{q,k} := \sum_{j=(1-q)k}^{k-1} Z_j = S_{1,k} - S_{1,(1-q)k}$ is the q -suffix cumulative sum (cumulative sum of the last qk terms) of the sequence Z_k . Then using (4.20), with probability at least $1 - \delta$,

$$\left\| \frac{S_{q,k}}{k} \right\| \leq \left\| \frac{S_{1,k}}{k} \right\| + \left\| \frac{S_{1,(1-q)k}}{k} \right\| \leq \frac{2t_\delta}{k} \quad (4.22)$$

Plugging this high probability bound into (4.21), we conclude.

(iii) By Lemma A.1, we have $\max_{1 \leq i < m} \|x_{i-1}^k - x^*\| = \mathcal{O}(\alpha_k)$. Therefore,

$$\begin{aligned} \|\nabla f_{\sigma_k(i)}(x_{i-1}^k) - \nabla f_{\sigma_k(i)}(x^*)\| &= \|P_{\sigma_k(i)}(x_{i-1}^k - x^*)\| \\ &\leq L \|x_{i-1}^k - x^*\| = \mathcal{O}(\alpha_k) \end{aligned} \quad (4.23)$$

for any $i = 1, 2, \dots, m$. As a consequence,

$$\begin{aligned} \hat{b}_{q,k} &= -\bar{\alpha}_{q,k} H_*^{-1} \sum_{i=1}^m P_{\sigma_k(i)} \left(\nabla f_{\sigma_k(i)}(x^*) + \mathcal{O}(\alpha_k) \right) \\ &= -\bar{\alpha}_{q,k} H_*^{-1} \sum_{j=1}^m P_j \nabla f_j(x^*) + \mathcal{O}(\alpha_k^2) \\ &= b_{q,k} + \mathcal{O}(\alpha_k^2) \end{aligned}$$

where in the second equality we use the fact that $\bar{\alpha}_{q,k} = \mathcal{O}(\alpha_k)$.

□

Part (i) of Theorem 4.3 shows that averaged RR iterates converge (if normalized properly with the stepsize) to a fixed vector that is non-zero in general. The magnitude of this (limit vector) drift depends on the averaging parameter q , the stepsize parameters s and R and also on μ_* which itself is a function of the steepness and the curvature (first and second derivatives) of the component functions at the optimal solution. The following toy example illustrates this.

EXAMPLE 4.1. Consider problem (1.1) with two quadratic functions in dimension one:

$$f_1(x) = \frac{1}{2}(x-1)^2, \quad f_2(x) = \frac{1}{2}(x+1)^2 + \frac{x^2}{2}$$

where $f(x) = \frac{3}{2}x^2 + 1$ and $x^* = 0$. The RR iterations $\{x_0^k\}$ become

$$x_0^{k+1} = x_0^k - \alpha_k(\nabla f(x_0^k) + E_k), \quad E_k = \alpha_k v(\sigma_k) - 2\alpha_k x_0^k, \quad (4.24)$$

where

$$v(\sigma_k) = \begin{cases} +2 & \text{with probability } 1/2, \quad \text{for } \sigma_k = \{1, 2\}, \\ -1 & \text{with probability } 1/2, \quad \text{for } \sigma_k = \{2, 1\}. \end{cases}$$

For this example, $P_1 = 1$, $P_2 = 2$, $H_* = 3$, $\nabla f_1(x^*) = -1$, $\nabla f_2(x^*) = 1$ and $\theta_* = 1$. Let $\alpha_k = 1/k^{3/4}$. Then $R = 1$, $s = 3/4$ and it follows from part (i) of Theorem 4.3,

$$\lim_{k \rightarrow \infty} k^s(\bar{x}_{q,k} - x^*) = -a_q(s)H_*^{-1}\theta_* = -\frac{1 - (1-q)^{1/4}}{q} \frac{1}{3}.$$

In contrast, SGD starting from the same initial point x_0^1 leads to the iterations

$$x_0^{k+1} = x_0^k - \alpha_k(\nabla f(x_0^k) + \bar{E}_k), \quad (4.25)$$

where the gradient error is given by

$$\bar{E}_k = \begin{cases} -2 - x_0^k - \alpha_k(x_0^k - 1) & \text{with probability } 1/4, \quad \text{for } \sigma_k = \{1, 1\}, \\ 2 + x_0^k - 2\alpha_k(2x_0^k + 1) & \text{with probability } 1/4, \quad \text{for } \sigma_k = \{2, 2\}, \\ E_k & \text{with probability } 1/4, \quad \text{for } \sigma_k = \{1, 2\}, \\ E_k & \text{with probability } 1/4, \quad \text{for } \sigma_k = \{2, 1\}, \end{cases}$$

which takes into the account that component functions can be sampled more than once in a cycle.

We next compare SGD and RR, focusing on what are the properties that makes RR achieve a faster rate than SGD. This is also important to gain insight for other algorithms where without-replacement sampling can outperform with-replacement sampling.

REMARK 4.2. In Example 4.1, SGD has a gradient error \bar{E}_k that has $\Theta(1)$ variance whereas the gradient errors in RR are $E_k = \mathcal{O}(\alpha_k)$ with a smaller variance $\mathcal{O}(\alpha_k^2)$. This behavior is not specific to this example but is typical (see Lemma A.3). A smaller variance in gradient errors leads to more accurate direction of descent and is the main reason behind the faster convergence of RR compared to SGD.

Note that the expectation of E_k is not zero in general for finite k and its distribution is not symmetric with respect to the origin. However, E_k goes to zero as the stepsize gets

smaller. This biasedness of E_k is also reflected in Theorem (4.3) that shows that the distribution of the approximation error $\bar{x}_{q,k} - x^*$ is not symmetric around zero in general (as $\theta_* \neq 0$ in general). This is in contrast to SGD with averaging which leads to an approximation error that converges to a centered normal distribution (see [24, Theorem 1]).

5. Extension to smooth component functions. Extending our results to more general smooth functions requires achieving similar bounds for the cycle gradient errors which depend on the gradients and Hessian matrices of the component functions along the inner iterates. In order to be able to control the change of gradients and Hessian matrices along the iterates, we introduce the following assumption which has also been used to analyze SGD [21].

ASSUMPTION 5.1. *The functions f_i are convex on \mathbb{R}^n and has Lipschitz continuous second derivatives, i.e. there exists a constant U_i such that*

$$\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\| \leq U_i \|x - y\|, \quad \forall x, y \in \mathbb{R}^n,$$

for $i = 1, 2, \dots, m$.

Under this assumption, by the triangle inequality, $\nabla^2 f(\cdot)$ is also Lipschitz with constant

$$U := \sum_{i=1}^m U_i. \tag{5.1}$$

When the component functions are quadratics, we have the special case with $U = U_i = 0$. We will now see how this assumption makes it possible to control the change of gradients of the component functions. Smooth functions f with Lipschitz Hessians are quadratic-like in the sense that the first-order Taylor approximation to the gradient of f is almost affine (with a quadratic term controlled by U) satisfying

$$\nabla f(x) = \nabla f(x^*) + H_*(x - x^*) + \eta, \quad \|\eta\| \leq \frac{U}{2} \|x - x^*\|^2, \quad \forall x. \tag{5.2}$$

(see e.g. [17, Section 1.3]) The analysis of Theorem 4.3 (and Lemma A.3 it builds upon) considers the $U = 0$ case (see e.g. (4.9) and (4.23)) applying a first-order Taylor approximation to the gradient of the component functions at $x = x_0^k$ satisfying $\|x - x^*\| = \mathcal{O}(\alpha_k)$. Therefore, when $U \neq 0$, an extra correction term $\eta = \mathcal{O}(\alpha_k^2)$ needs to be added to the analysis. However, we show in the next theorem that this correction term does not cause a slow down in the convergence rate compared to the quadratic case because the q -suffix averages of this $\mathcal{O}(\alpha_k^2)$ correction term decays like $\mathcal{O}(1/k)^2$.

We will also need one more technical assumption that appeared in a number of papers in the literature for analyzing incremental methods to rule out the case that the iterates

²This is due to the fact that the sequence α_k^2 is summable when $s > 1/2$.

diverge to infinity. In particular, this assumption is made in [18] for generalizing Theorem 4.1 on the rate of deterministic IG from quadratic functions to general smooth functions which we will be referring to.

ASSUMPTION 5.2. *Iterates $\{x_j^k\}_{j,k}$ generated are uniformly bounded, i.e. there exists a non-empty compact Euclidean ball $\mathcal{X} \subset \mathbb{R}^n$ that contains all the iterates a.s.*³

Equipped with these two assumptions, all the results of Theorem 4.3) extend naturally with minor modifications. In particular, P_i (which is a constant Hessian matrix in the setting of Theorem 4.3) needs to be replaced by $\nabla^2 f_i(x^*)$ or $\nabla^2 f_i(x_{i-1}^k)$ depending on the context.

THEOREM 5.1. *Consider the RR iterations given by (1.4) with stepsize $\alpha_k = \frac{R}{(k+1)^s}$ where $R > 0$ and $s \in (\frac{1}{2}, 1)$. Suppose that Assumptions 3.1, 5.1 and 5.2 hold. Then the following statements are true:*

(i) *For any $0 < q \leq 1$,*

$$\lim_{k \rightarrow \infty} k^s (\bar{x}_{q,k} - x^*) = a_q(s) H_*^{-1} \theta_* \quad a.s.$$

where $H_* = \nabla^2 f(x^*)$ is the Hessian matrix at the optimal solution,

$$a_q(s) = -\frac{1 - (1-q)^{1-s}}{q} R \quad \text{and} \quad \theta_* = \frac{1}{2} \sum_{i=1}^m \nabla^2 f_i(x^*) \nabla f_i(x^*). \quad (5.3)$$

(ii) *With probability at least $1 - \delta$, we have*

$$\bar{x}_{q,k} - x^* = r_{q,k} + \mathcal{O}\left(\frac{\sqrt{\log(1/\delta)}}{k}\right) + \begin{cases} \mathcal{O}(\frac{\log k}{k}) & \text{if } q = 1 \\ \mathcal{O}(\frac{1}{k}) & \text{if } 0 < q < 1, \end{cases}$$

where

$$r_{q,k} = -\bar{\alpha}_{q,k} H_*^{-1} \theta_* \quad (5.4)$$

is deterministic. The constants hidden by $\mathcal{O}(\cdot)$ depend only on G_*, L, m, R, c, q, s and U .

(iii) *Let*

$$\hat{r}_{q,k} = -\bar{\alpha}_{q,k} \left[\sum_{i=1}^m \nabla^2 f_{\sigma_k(i)}(x_{i-1}^k) \right]^{-1} \sum_{i=1}^m \nabla^2 f_{\sigma_k(i)}(x_{i-1}^k) \nabla f_{\sigma_k(i)}(x_{i-1}^k).$$

Then,

$$\hat{r}_{q,k} = r_{q,k} + \mathcal{O}(\alpha_k^2).$$

³Note that if this assumption holds and if f_i is three-times continuously differentiable on the compact set \mathcal{X} , then the third-order derivatives are bounded and Assumption 5.1 holds.

It follows from part (ii) that with probability at least $1 - \delta$,

$$(\bar{x}_{q,k} - \hat{r}_{q,k}) - x^* = \mathcal{O}\left(\frac{\sqrt{\log(1/\delta)}}{k}\right) + \begin{cases} \mathcal{O}(\frac{\log k}{k}) & \text{if } q = 1 \\ \mathcal{O}(\frac{1}{k}) & \text{if } 0 < q < 1. \end{cases}$$

Proof. The proof of Theorem 4.3 applies almost word by word except that the Taylor approximation for the gradients of the component functions will have an extra term compared to the proof of Theorem 4.3 (see also (5.2)). Also, instead of Lemmas A.2 and A.3 that apply to only quadratic functions, their extensions Lemmas B.2 and B.3 given in the appendix are used in the proof. Besides these changes, for the sake of completeness we also give an overview of the main modifications required for each part of the proof:

- (i) The expression (4.9) for the gradient should be modified to include an extra error term η_j of the form

$$\nabla f(x_0^j) = H_*(x_0^j - x^*) + \eta_j, \quad \|\eta_j\| \leq \frac{U}{2} \|x_0^j - x^*\|^2 \quad (5.5)$$

By Lemma B.1, $\sum_j \eta_j \leq \frac{U}{2} \|x_0^j - x^*\|^2 = \mathcal{O}(\alpha_j^2)$ therefore the sequence η_j is summable and if averaged decays like $\mathcal{O}(1/k)$ without degrading the convergence rate except possibly the constants hidden by $\mathcal{O}(\cdot)$.

- (ii) The same proof applies by invoking Lemma B.3 in lieu of Lemma A.3.
- (iii) Instead of Lemma A.1, we use Lemma B.1. The expression (4.23) on the difference of gradients needs to be adjusted as

$$\|\nabla f_{\sigma_k(i)}(x_{i-1}^k) - \nabla f_{\sigma_k(i)}(x^*) - \nabla^2 f_{\sigma_k(i)}(x^*)(x_{i-1}^k - x^*)\| \leq \frac{U}{2} \|x_i^k - x^*\|^2. \quad (5.6)$$

The righthand side is still $\mathcal{O}(\alpha_k^2)$ by an application of Lemma B.1 therefore the rest of the proof applies.

□

6. An accelerated RR algorithm. Part (iii) of Theorem 5.1 (see also part (iii) of Theorem 4.3) shows that if the bias term $\hat{r}_{q,k}$ is subtracted from the q -suffix averaged RS iterates, then the distance to the optimal solution of the q -suffix averaged iterates becomes on the order of $\mathcal{O}(1/k)$ with high probability. By strong convexity, this translates into a rate of $\mathcal{O}(1/k^2)$ in the suboptimality of the objective values. We call this “subtraction operation”, *bias removal*. Algorithm 6 which we call *BIRR* (Bias Removed Random Reshuffling), describes how this can be done efficiently. Note that it suffices to do this subtraction only once at the last cycle.

The bias removal of the BIRR algorithm requires an $n \times n$ matrix inversion which requires $\approx n^3$ arithmetic operations (if there is more structure on the Hessian of f_i such as low-rankness or sparsity this could be improved to $\approx n^2$ with Woodbury-Morrison type

Algorithm 1 Bias Removed Random Reshuffling with Suffix Averaging (BIRR)

Input: Initial point $x_0^0 \in \mathbb{R}^n$, number of cycles $K \in \mathbb{N}$, suffix averaging parameter $q \in (0, 1]$, stepsize parameters $R > 0$ and $s \in (1/2, 1)$.

Initialization: $\bar{x}_{1,0} = 0 \in \mathbb{R}^n$, $\hat{\mu}_0 = 0 \in \mathbb{R}^n$, $\bar{\alpha}_{1,0} = 0 \in \mathbb{R}$, $\hat{H}_0 = 0 \in \mathbb{R}^{n \times n}$.

1. For each cycle $k = 0, 1, 2, \dots, K - 1$:

(a) Inner iteration.

Pick a permutation σ_k of $\{1, \dots, m\}$ uniformly at random.

For $i = 1, 2, \dots, m$:

Compute x_i^k by:

$$x_i^k = x_{i-1}^k - \alpha_k \nabla f_{\sigma_k(i)}(x_{i-1}^k), \quad \alpha_k = \frac{R}{(k+1)^s}.$$

// Precompute for the bias estimation only for the last cycle

If $k = K - 1$, compute $\hat{\mu}_i$ and \hat{H}_i by :

$$\begin{aligned} \hat{\mu}_i &= \hat{\mu}_{i-1} + \nabla^2 f_{\sigma_k(i)}(x_{i-1}^k) \nabla f_{\sigma_k(i)}(x_{i-1}^k) \\ \hat{H}_i &= \hat{H}_{i-1} + \nabla^2 f_{\sigma_k(i)}(x_{i-1}^k) \end{aligned}$$

Set outer iterate: $x_0^{k+1} = x_m^k$.

(b) Update the simple average of the iterates and the stepsize:

$$\begin{aligned} \bar{x}_{1,k+1} &= \frac{k}{k+1} \bar{x}_{1,k} + \frac{1}{k+1} x_0^k \\ \bar{\alpha}_{1,k+1} &= \frac{k}{k+1} \bar{\alpha}_{1,k} + \frac{1}{k+1} \alpha_k \end{aligned}$$

2. If $q \in (0, 1)$, compute q -suffix averages from the simple averages:

$$\bar{x}_{q,K} = \frac{\bar{x}_{1,K} - q\bar{x}_{1,(1-q)K}}{1-q}, \quad \bar{\alpha}_{q,K} = \frac{\bar{\alpha}_{1,K} - q\bar{\alpha}_{1,(1-q)K}}{1-q}.$$

3. Estimate the bias by the formula (4.7) : $\hat{b}_{q,K} = -\bar{\alpha}_{q,K} \hat{H}_m^{-1} \hat{\mu}_m$ in the last cycle.

Output: $\bar{x}_{q,K} - \hat{b}_{q,K}$.

of approach), but accelerates the convergence with high-probability. For small or moderate n , this could be done efficiently and incrementally processing the functions one at a time; however for large n this may be impractical or infeasible depending on the application in which case the bias estimation and removal step can be skipped.

7. Conclusion. We analyzed the random reshuffling (RR) method for minimizing a finite sum of convex component functions. When the objective function is strongly convex and the component functions are smooth, averaged RR iterates converge at rate $\sim 1/k^s$ to the optimal solution almost surely (which translates into a rate of $1/k^{2s}$ in the suboptimality

of the objective value) for a diminishing stepsize $\alpha_k = \Theta(1/k^s)$. This is faster than SGD's $\Omega(\frac{1}{k})$ min-max rate. Viewing RR as a gradient descent method with random gradient errors, this result builds on first showing that gradient errors $E_k = \mathcal{O}(\alpha_k)$ and then relating the gradient error sequence to an i.i.d sequence to which martingale theory is applicable. Note that the gradient errors in SGD are larger with a $\mathcal{O}(1)$ variance, which leads to a less accurate gradient descent direction. Beyond RR and SGD comparison, these results also give insight into the fast convergence properties of without-replacement sampling strategies compared to with-replacement sampling strategies.

After characterizing the convergence rate of RR, we look into second-order terms in the asymptotic expansion of the averaged RR iterates and obtain high probability bounds. We use this bounds to develop a new method that can accelerate the convergence rate of RR to $\mathcal{O}(\frac{1}{k^2})$ with high probability.

Appendix A. Technical lemmas for the proof of Theorem 4.3.

The first lemma is on characterizing what is the worst-case distance of the all the inner iterates of RR to the optimal solution x^* . This quantity we want to upper bound is a random variable, but the upper bounds we obtain are deterministic holding for every random path. This lemma is based on Corollary 4.2 and uses the fact that the distance between the inner iterates are on the order of the stepsize.

LEMMA A.1. *Under the conditions of Theorem 4.3 we have*

$$\max_{0 \leq i < m} \|x_i^k - x^*\| = \mathcal{O}\left(\frac{1}{k^s}\right).$$

where $\mathcal{O}(\cdot)$ hides a constant that depends only on G_*, L, m, c and R .

Proof. By Corollary 4.2,

$$\|x_0^k - x^*\| = \mathcal{O}\left(\frac{1}{k^s}\right). \quad (\text{A.1})$$

where $\mathcal{O}(\cdot)$ hides a constant that depends only on G_*, L, m, R and c . We have also for any $0 \leq i < m$ and $k \geq 0$,

$$\begin{aligned} \|x_i^k - x^*\| &\leq \|x_0^k - x^*\| + \|x_i^k - x_0^k\| = \|x_0^k - x^*\| + i\alpha_k \max_{\ell=1, \dots, i} \|\nabla f_{\sigma_k(\ell)}(x_{\ell-1}^k)\| \\ &\leq \|x_0^k - x^*\| + (m-1) \frac{R}{(k+1)^s} (G_* + \max_{\ell=1, \dots, i} \|\nabla f_{\sigma_k(\ell)}(x_{\ell-1}^k) - \nabla f_{\sigma_k(\ell)}(x^*)\|) \\ &\leq \|x_0^k - x^*\| + (m-1) \frac{R}{(k+1)^s} (G_* + L \max_{\ell=1, \dots, i} \|x_{\ell-1}^k - x^*\|). \end{aligned}$$

where we used the L -Lipschitzness of the gradient of f where L is given by 4.2. Using Equation (A.1) and applying this inequality inductively for $i = 0, 1, 2, \dots, m-1$ we conclude.

□

The second lemma is on characterizing how fast on average the outer iterates move (if normalized by the stepsize) after a cycle of the RR algorithm. This is clearly related to

the magnitude of the gradients seen by the iterates and is fundamental for establishing the convergence rate of the averaged RR iterates in Theorem 4.3.

LEMMA A.2. *Under the conditions of Theorem 4.3, consider the sequence*

$$I_{q,k} = \frac{\sum_{j=(1-q)k}^{k-1} (x_0^j - x_0^{j+1}) \alpha_j^{-1}}{qk}, \quad 0 < q \leq 1. \quad (\text{A.2})$$

Then,

$$I_{q,k} = \begin{cases} \mathcal{O}(\frac{\log k}{k}) & \text{if } q = 1, \\ \mathcal{O}(\frac{1}{k}) & \text{if } 0 < q < 1. \end{cases} \quad (\text{A.3})$$

In the former case, $\mathcal{O}(\cdot)$ hides a constant that depends only on G_*, L, m, c, R, s, q and dist_0 . In the latter case, the same dependency on the constants occurs except that the dependency on dist_0 can be removed.

Proof. It follows from integration by parts that for any $\ell < k$,

$$-\sum_{j=\ell}^{k-1} (x_0^j - x_0^{j+1}) \alpha_j^{-1} = \alpha_k^{-1} (x_0^k - x^*) - \alpha_\ell^{-1} (x_0^\ell - x^*) - \sum_{j=\ell}^{k-1} (x_0^{j+1} - x^*) (\alpha_{j+1}^{-1} - \alpha_j^{-1}). \quad (\text{A.4})$$

Next, we investigate the asymptotic behavior of the terms on the right-hand side. A consequence of Corollary 4.2 and Equation (4.4) is that

$$\alpha_k^{-1} \|x_0^k - x^*\| = \frac{(k+1)^s}{R} \|x_0^k - x^*\| \leq \frac{LmG_*}{c} + o(1) = \mathcal{O}(1) \quad (\text{A.5})$$

and therefore

$$\begin{aligned} |\alpha_{k+1}^{-1} - \alpha_k^{-1}| \|x_0^k - x^*\| &= \frac{(k+2)^s - (k+1)^s}{(k+1)^s} \alpha_k^{-1} \|x_0^k - x^*\| \\ &= \left(\left(1 + \frac{1}{k+1}\right)^s - 1 \right) \alpha_k^{-1} \|x_0^k - x^*\| \leq \frac{s}{k+1} \alpha_k^{-1} \|x_0^k - x^*\| \\ &\leq \frac{sLmG_*}{c} \frac{1}{k+1} + o\left(\frac{1}{k+1}\right) = \mathcal{O}\left(\frac{1}{k+1}\right) \end{aligned}$$

where $\mathcal{O}(\cdot)$ hides a constant that depend only on L, G_*, c, m and s . Then, setting $\ell = (1-q)k$ in (A.4), it follows that

$$\begin{aligned} \left\| \sum_{j=\ell}^{k-1} (x_0^j - x_0^{j+1}) \alpha_j^{-1} \right\| &\leq \|\alpha_k^{-1} (x_0^k - x^*)\| + \|\alpha_{(1-q)k}^{-1} (x_0^{(1-q)k} - x^*)\| \\ &\quad + \sum_{j=(1-q)k}^{k-1} \|x_0^{j+1} - x^*\| |\alpha_{j+1}^{-1} - \alpha_j^{-1}|. \\ &= \mathcal{O}(1) + \|\alpha_{(1-q)k}^{-1} (x_0^{(1-q)k} - x^*)\| + \mathcal{O}\left(\sum_{j=(1-q)k}^{k-1} \frac{1}{j+1}\right). \end{aligned} \quad (\text{A.6})$$

We also have

$$\|\alpha_{(1-q)k}^{-1} (x_0^{(1-q)k} - x^*)\| = \begin{cases} \alpha_0^{-1} \text{dist}_0 & \text{if } q = 1, \\ \mathcal{O}(1) & \text{if } 0 < q < 1, \end{cases} \quad (\text{A.8})$$

where the second part follows from (A.5) with similar constants for the $\mathcal{O}(\cdot)$ term. As the sequence $\frac{1}{j+1}$ is monotonically decreasing, for any $k > 0$ we have the bounds

$$\sum_{j=(1-q)k}^{k-1} \frac{1}{j+1} \leq 1 + \int_{(1-q)k}^k \frac{1}{x+1} dx = \begin{cases} \log k + 1 & \text{if } q = 1 \\ \log(\frac{1}{1-q}) + 1 & \text{if } 0 < q < 1. \end{cases} \quad (\text{A.9})$$

Note that when $q = 1$ this bound grows with k logarithmically whereas for $q < 1$ it does not grow with k . Then, combining (A.7), (A.8) and (A.9) we obtain

$$\|I_{q,k}\| \leq \frac{\|\sum_{j=\ell}^{k-1} (x_0^j - x_0^{j+1}) \alpha_j^{-1}\|}{qk} = \begin{cases} \mathcal{O}(\frac{\log k}{k}) & \text{if } q = 1 \\ \mathcal{O}(\frac{1}{k}) & \text{if } 0 < q < 1 \end{cases}$$

as desired which completes the proof. \square

LEMMA A.3. *Under the conditions of Theorem 4.3, the following statements are true:*

(i) *We have*

$$E_k = \alpha_k v(\sigma_k) + \mathcal{O}(\alpha_k^2), \quad k \geq 0, \quad (\text{A.10})$$

where E_k is the gradient error defined by (3.2), $\mathcal{O}(\cdot)$ hides a constant that depends only on G_* , L , m , R and c and

$$v(\sigma_k) = - \sum_{i=1}^m P_{\sigma_k(i)} \sum_{\ell=1}^{i-1} \nabla f_{\sigma_k(\ell)}(x^*). \quad (\text{A.11})$$

(ii) *It holds that*

$$\|v(\sigma_k)\| = M_{\sigma_k} \leq LmG_* \quad (\text{A.12})$$

where M_{σ_k} is as in Theorem 4.1, $L = \sum_i \|P_i\|$ and G_* is defined by (4.3). Furthermore,

$$\mathbb{E}v(\sigma_k) = \mu_* = \frac{1}{2} \sum_{i=1}^m P_i \nabla f_i(x^*). \quad (\text{A.13})$$

(iii) *For any $0 < q \leq 1$, $\lim_{k \rightarrow \infty} Y_{q,k} = \mu_*$ a.s. where*

$$Y_{q,k} = \frac{\sum_{i=(1-q)k}^{k-1} E_j}{\sum_{j=(1-q)k}^{k-1} \alpha_j}. \quad (\text{A.14})$$

Proof.

(i) As component functions are quadratics, (3.2) becomes

$$E_k = \sum_{i=1}^m P_{\sigma_k(i)} (x_{i-1}^k - x_0^k) = - \sum_{i=1}^m P_{\sigma_k(i)} \alpha_k \sum_{\ell=1}^{i-1} \nabla f_{\sigma_k(\ell)}(x_{\ell-1}^k).$$

where we can substitute

$$\nabla f_{\sigma_k(\ell)}(x_{\ell-1}^k) = \nabla f_{\sigma_k(\ell)}(x^*) + P_{\sigma_k(\ell)}(x_{\ell-1}^k - x^*). \quad (\text{A.15})$$

Then an application of Lemma A.1 proves directly the desired result.

(ii) For the first part, the inequality (A.12) is just a consequence of the triangle inequality applied to the definition (A.11) with $L_i = \|P_i\|$ and $L = \sum_{i=1}^m L_i$. For the second part, note that for any $i \neq \ell$, the joint distribution of $(\sigma_k(i), \sigma_k(\ell))$ is uniform over the set of all (ordered) pairs from $\{1, 2, \dots, m\}$. Therefore, for any $i \neq \ell$,

$$\begin{aligned}\mathbb{E}[P_{\sigma_k(i)} \nabla f_{\sigma_k(\ell)}(x^*)] &= \sum_{i=1}^m \sum_{i \neq j, j=1}^m \frac{P_i \nabla f_j(x^*)}{m(m-1)} \\ &= \frac{\sum_{i=1}^m P_i \sum_{j=1}^m \nabla f_j(x^*) - \sum_{j=1}^m P_j \nabla f_j(x^*)}{m(m-1)} \\ &= -\frac{\sum_{j=1}^m P_j \nabla f_j(x^*)}{m(m-1)}\end{aligned}$$

where we used the fact that $\nabla f(x^*) = \sum_{j=1}^m \nabla f_j(x^*) = 0$ by the first order optimality condition. Then, by taking the expectation of (A.11), we obtain

$$\begin{aligned}\mathbb{E}v(k) &= -\sum_{i=1}^m \sum_{\ell=0}^{i-1} \mathbb{E}[P_{\sigma_k(i)} \nabla f_{\sigma_k(\ell)}(x^*)] \\ &= -\sum_{i=1}^m \sum_{\ell=0}^{i-1} -\frac{\sum_{j=1}^m P_j \nabla f_j(x^*)}{m(m-1)} \\ &= \frac{\sum_{j=1}^m P_j \nabla f_j(x^*)}{2}.\end{aligned}$$

which completes the proof.

(iii) We introduce the normalized gradient error sequence $Y_j = E_j/\alpha_j$. By part (i), $Y_j = v(\sigma_j) + \mathcal{O}(\alpha_j)$ where $v(\sigma_j)$ is a sequence of i.i.d. variables. By the strong law of large numbers, we have

$$\lim_{k \rightarrow \infty} \frac{\sum_{j=0}^{k-1} v(\sigma_j)}{k} = \mathbb{E}v(\sigma_j) = \mu_* \quad \text{a.s.} \quad (\text{A.16})$$

where the last equality follows from part (ii). Therefore,

$$\lim_{k \rightarrow \infty} \frac{\sum_{j=0}^{k-1} Y_j}{k} = \lim_{k \rightarrow \infty} \left(\frac{\sum_{j=0}^{k-1} v(\sigma_j)}{k} + \frac{\sum_{j=0}^{k-1} \mathcal{O}(\alpha_j)}{k} \right) = \mu_* \quad \text{a.s.}$$

where we used the fact that the second term is negligible as $\sum_{j=0}^{k-1} \alpha_j/k = \mathcal{O}(k^{-s}) \rightarrow 0$. As the average of the sequence Y_j converges almost surely, one can show that this implies almost sure convergence of a weighted average of the sequence Y_j as well as long as weights satisfy certain conditions as $k \rightarrow \infty$. In particular, as the sequence $\{\alpha_j\}$ is monotonically decreasing and is non-summable, by [14, Theorem 1],

$$\lim_{k \rightarrow \infty} Y_{1,k} = \lim_{k \rightarrow \infty} \frac{\sum_{j=0}^{k-1} \alpha_j Y_j}{\sum_{j=0}^{k-1} \alpha_j} = \lim_{k \rightarrow \infty} \frac{\sum_{j=0}^{k-1} E_j}{\sum_{j=0}^{k-1} \alpha_j} = \mu_* \quad \text{a.s.} \quad (\text{A.17})$$

This completes the proof for $q = 1$. For $0 < q < 1$, by the definition of $Y_{q,k}$, we can write

$$Y_{1,k} = (1 - w_k)Y_{q,k} + w_k Y_{1,(1-q)k}$$

where the non-negative weights w_k satisfy

$$w_k = \frac{\sum_{j=0}^{(1-q)k-1} \alpha_j}{\sum_{j=0}^{k-1} \alpha_j} \xrightarrow{k \rightarrow \infty} (1-q)^{1-s} < 1.$$

As both $Y_{1,k}$ and $Y_{1,(1-q)k}$ go to μ_* a.s. by (A.17), it follows that

$$\lim_{k \rightarrow \infty} Y_{q,k} = \lim_{k \rightarrow \infty} \frac{Y_{1,k} - w_k Y_{1,(1-q)k}}{1 - w_k} = \mu_* \quad \text{a.s.}$$

too for any $0 < q < 1$. This completes the proof.

□

Appendix B. Technical Lemmas for the proof of Theorem 5.1. We first state the following result from [18] which extends Corollary 4.2 from quadratics to smooth functions.

COROLLARY B.1. *Under the setting of Theorem 5.1, then*

$$\text{dist}_k \leq \frac{RM}{c} \frac{1}{k^s} + o\left(\frac{1}{k^s}\right),$$

where the right-hand side is a deterministic sequence and

$$M = LmG_* \quad \text{where} \quad G_* = \sup_{1 \leq i \leq m} \|\nabla f_i(x^*)\|.$$

Proof. The proof of Corollary 4.2 is based on Theorem 4.1 from [18]. This theorem admit an extension to smooth functions with Lipschitz gradients to (see [18]), therefore by the same reasoning along the lines of Corollary 4.2 the result follows. □

LEMMA B.1. *Under the conditions of Theorem 5.1, all the conclusions of Lemma A.1 remain valid.*

Proof. The proof of Lemma A.1 applies identically except that instead of Corollary 4.2 we use its extension Corollary (B.1). □

LEMMA B.2. *Under the conditions of Theorem 5.1, all the conclusions of Lemma A.2 remain valid.*

Proof. The proof of Lemma A.2 applies identically with the only difference that the bound on $\|x_0^k - x^*\|$ is obtained from Corollary B.1 instead of Corollary 4.2. □

LEMMA B.3. *Under the conditions of Theorem 5.1, the following statements are true:*

(i) *We have*

$$E_k = \alpha_k \bar{v}(\sigma_k) + \mathcal{O}(\alpha_k^2), \quad k \geq 0, \tag{B.1}$$

where $\mathcal{O}(\cdot)$ hides a constant that depends only on G_*, L, m, R, c and U and

$$\bar{v}(\sigma_k) = - \sum_{i=0}^{m-1} \nabla^2 f_{\sigma_k(i)}(x^*) \sum_{\ell=0}^{i-1} \nabla f_{\sigma_k(\ell)}(x^*).$$

(ii) It holds that

$$\|\bar{v}(\sigma_k)\| \leq LmG_* \quad (\text{B.2})$$

where

$$\theta_* := \mathbb{E}\bar{v}(\sigma_k) = \sum_{i=1}^m \nabla^2 f_i(x^*) \nabla f_i(x^*) / 2. \quad (\text{B.3})$$

(iii) For any $0 < q \leq 1$, $\lim_{k \rightarrow \infty} Y_{q,k} = \theta_*$ with probability one where

$$Y_{q,k} = \frac{\sum_{j=(1-q)k}^{k-1} E_j}{\sum_{j=(1-q)k}^{k-1} \alpha_j}. \quad (\text{B.4})$$

Proof. For part (i), first we express E_k using the Taylor expansion and the Hessian Lipschitzness as

$$\begin{aligned} E_k &= \sum_{i=1}^m \left(\nabla^2 f_{\sigma_k(i)}(x_0^k) \right) (x_{i-1}^k - x_0^k) + \mathcal{O}(U \|x_{i-1}^k - x_0^k\|^2). \\ &= - \sum_{i=1}^m \left(\nabla^2 f_{\sigma_k(i)}(x_0^k) \right) (x_{i-1}^k - x_0^k) + \mathcal{O} \left(\alpha_k^2 U \left\| \sum_{\ell=1}^{i-1} \nabla f_{\sigma_k(\ell)}(x_{\ell-1}^k) \right\| \right) \end{aligned}$$

By Lemma B.1, we have $\|x_\ell^k - x^*\| = \mathcal{O}(\alpha^k)$ with probability one. Then, by the gradient and Hessian Lipschitzness we can substitute above

$$\begin{aligned} \nabla f_{\sigma_k(\ell)}(x_{\ell-1}^k) &= \nabla f_{\sigma_k(\ell)}(x^*) + \mathcal{O}(\alpha^k), \\ \nabla^2 f_{\sigma_k(\ell)}(x_{\ell-1}^k) &= \nabla^2 f_{\sigma_k(\ell)}(x^*) + \mathcal{O}(\alpha^k). \end{aligned}$$

which implies directly Equation (B.1). The rest of the proof for parts (ii) and (iii) is similar to the proof of Lemma A.3 and is omitted. \square

REFERENCES

- [1] A. Agarwal, P.L. Bartlett, P. Ravikumar, and M.J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, May 2012.
- [2] D. Bertsekas. Incremental least squares methods and the extended kalman filter. *SIAM Journal on Optimization*, 6(3):807–822, 1996.
- [3] D. Bertsekas. A hybrid incremental gradient method for least squares. *SIAM Journal on Optimization*, 7:913–926, 1997.
- [4] D. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.
- [5] D. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. *Optimization for Machine Learning*, 2010:1–38, 2011.

- [6] D. Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, 2015.
- [7] L. Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science, Paris*, 2009.
- [8] L. Bottou. Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier and G. Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag HD, 2010.
- [9] L. Bottou. Stochastic Gradient Descent Tricks. In *Neural Networks: Tricks of the Trade*, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [10] L. Bottou and Y. Le Cun. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005.
- [11] L. Bottou. Stochastic gradient descent tricks. In Grgoire Montavon, GenevieveB. Orr, and Klaus-Robert Mller, editors, *Neural Networks: Tricks of the Trade*, volume 7700 of *Lecture Notes in Computer Science*, pages 421–436. Springer Berlin Heidelberg, 2012.
- [12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
- [13] K. L. Chung. On a Stochastic Approximation Method. *Annals of Mathematical Statistics*, 25(3):463–483, September 1954.
- [14] N. Etemadi. Convergence of weighted averages of random variables revisited. *Proceedings of the American Mathematical Society*, 134(9):2739–2744, 2006.
- [15] V Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 1968.
- [16] X. Feng, A. Kumar, B. Recht, and C. Ré. Towards a unified architecture for in-rdbms analytics. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 325–336. ACM, 2012.
- [17] N.I.M. Gould and S. Leyffer. An introduction to algorithms for nonlinear optimization. In J.F. Blowey, A.W. Craig, and T. Shardlow, editors, *Frontiers in Numerical Analysis*, Universitext, pages 109–197. Springer Berlin Heidelberg, 2003.
- [18] M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo. Convergence rate of incremental gradient and incremental Newton methods. *In Preparation*, 2015.
- [19] A. Israel, F. Krahmer, and R. Ward. An arithmetic-geometric mean inequality for products of three matrices. *Linear Algebra and its Applications*, 488:1 – 12, 2016.
- [20] H. J. Kushner and G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.
- [21] E. Moulines and F. R. Bach. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. *Advances in Neural Information Processing*, pages 451–459, 2011.
- [22] A. Nedić and A. Ozdaglar. On the rate of convergence of distributed subgradient methods for multi-agent optimization. In *Proceedings of IEEE CDC*, pages 4711–4716, 2007.
- [23] A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.
- [24] B. T. Polyak and A. B. Juditsky. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, July 2006.
- [25] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, Sep 2011.
- [26] S.S. Ram, A. Nedic, and V.V. Veeravalli. Stochastic incremental gradient descent for estimation in sensor networks. In *Signals, Systems and Computers, ACSSC 2007.*, pages 582–586, 2007.

- [27] B. Recht and C. Ré. Toward a noncommutative arithmetic-geometric mean inequality: Conjectures, case-studies, and consequences. *JMLR Workshop and Conference Proceedings*, 23:11.1–11.24, 2012.
- [28] B. Recht and C. Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.
- [29] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 693–701. Curran Associates, Inc., 2011.
- [30] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, 1951.
- [31] N. L. Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2663–2671. Curran Associates, Inc., 2012.
- [32] O Shamir. Open Problem: Is Averaging Needed for Strongly Convex Stochastic Gradient Descent? *COLT*, 2012.
- [33] J. Sohl-Dickstein, B. Poole, and S. Ganguli. Fast large-scale optimization by unifying stochastic gradient and quasi-Newton methods. In T. Jebara and E. P. Xing, editors, *ICML*, pages 604–612. JMLR Workshop and Conference Proceedings, 2014.
- [34] E.R. Sparks, A. Talwalkar, V. Smith, J. Kottalam, P. Xinghao, J. Gonzalez, M.J. Franklin, M.I Jordan, and T. Kraska. MLI: An API for distributed machine learning. In *IEEE 13th International Conference on Data Mining (ICDM)*, pages 1187–1192, 2013.
- [35] D. B. Yudin and A. Nemirovskii. Problem complexity and method efficiency in optimization, 1983.
- [36] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-first International Conference on Machine Learning, ICML*, pages 116–, New York, NY, USA, 2004. ACM.
- [37] T. Zhang. A note on the non-commutative arithmetic-geometric mean inequality. *arXiv preprint arXiv:1411.5058*, November 2014.